EA2022-analytical-V3.R

Edre MA

2022-03-15

```
#=======================
#Analytical statistics
#R Biostat Workshop IIUM
#Edre MA, DrPH
#=======================

#objective 3: To determine the factors contributing to hypertension
#we want to know first what contributes to systolic hypertension
```

```
#Comparing numerical values: parametric
```

```
install.packages('readr')
```

```
## Error in install.packages : Updating loaded packages
```

```
library(readr)
hstat <- read_csv("healthstatus6.csv")
```

```
## Rows: 153 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): sex, exercise, smoking
## dbl (14): id, age, wt, ht, sbp, dbp, hba1c, hcy, wt2, wt3, sbp2, sbp3, dbp2...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(hstat)

#if our IV is categorical with 2 categories
#example, we want to know if being male has any relationship with sbp
#independent sample t test

install.packages("car") #testing for homogeneity of variance
```

```
## Error in install.packages : Updating loaded packages
```

```r
library(car)
leveneTest(sbp ~ sex, data = hstat, center=mean)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##        Df F value Pr(>F)
## group   1  0.5477 0.4604
##       151
```

```r
t.test(sbp ~ sex, data = hstat)
```

```
##
##  Welch Two Sample t-test
##
## data:  sbp by sex
## t = 0.4972, df = 141.8, p-value = 0.6198
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.928972  3.225357
## sample estimates:
## mean in group Female   mean in group Male
##             132.6000             131.9518
```

```r
#we want to visualize the comparison

install.packages("ggplot2")
```

```
## Error in install.packages : Updating loaded packages
```

```r
library(ggplot2)
install.packages("ggpubr")
```

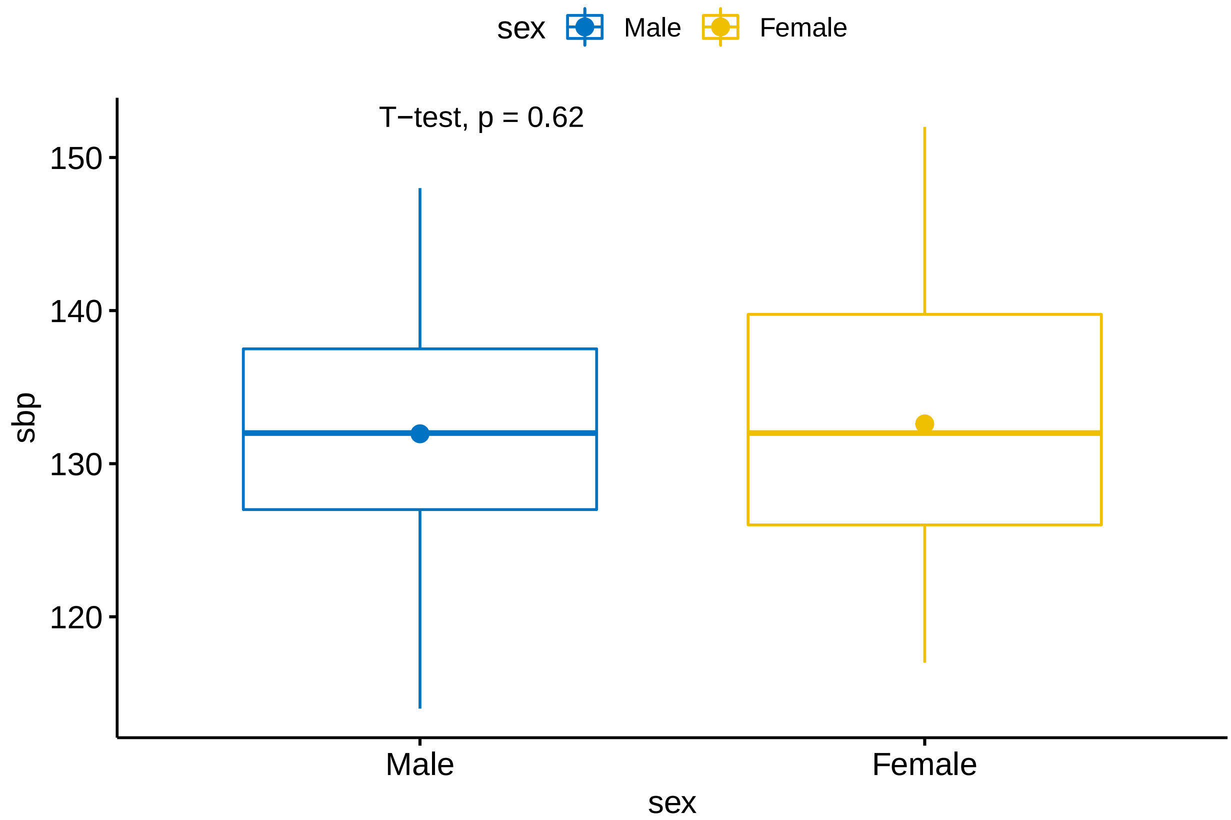
```
## Error in install.packages : Updating loaded packages
```

```r
library(ggpubr)
ggboxplot(hstat, x = "sex", y = "sbp",
          color = "sex",
          palette = "jco",
          add = "mean") +
          stat_compare_means(method = "t.test")
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

```
## Warning: `fun.ymin` is deprecated. Use `fun.min` instead.
```

```
## Warning: `fun.ymax` is deprecated. Use `fun.max` instead.
```




```
#now we know sex has no effect on sbp in our study
#we want to know now, does exercise have an effect (low,mod,high intensity)
#one way ANOVA

install.packages("psych")
```

```
## Error in install.packages : Updating loaded packages
```

```
library(psych)
describe.by(hstat$sbp, hstat$exercise)
```

```
## Warning: describe.by is deprecated. Please use the describeBy function
```

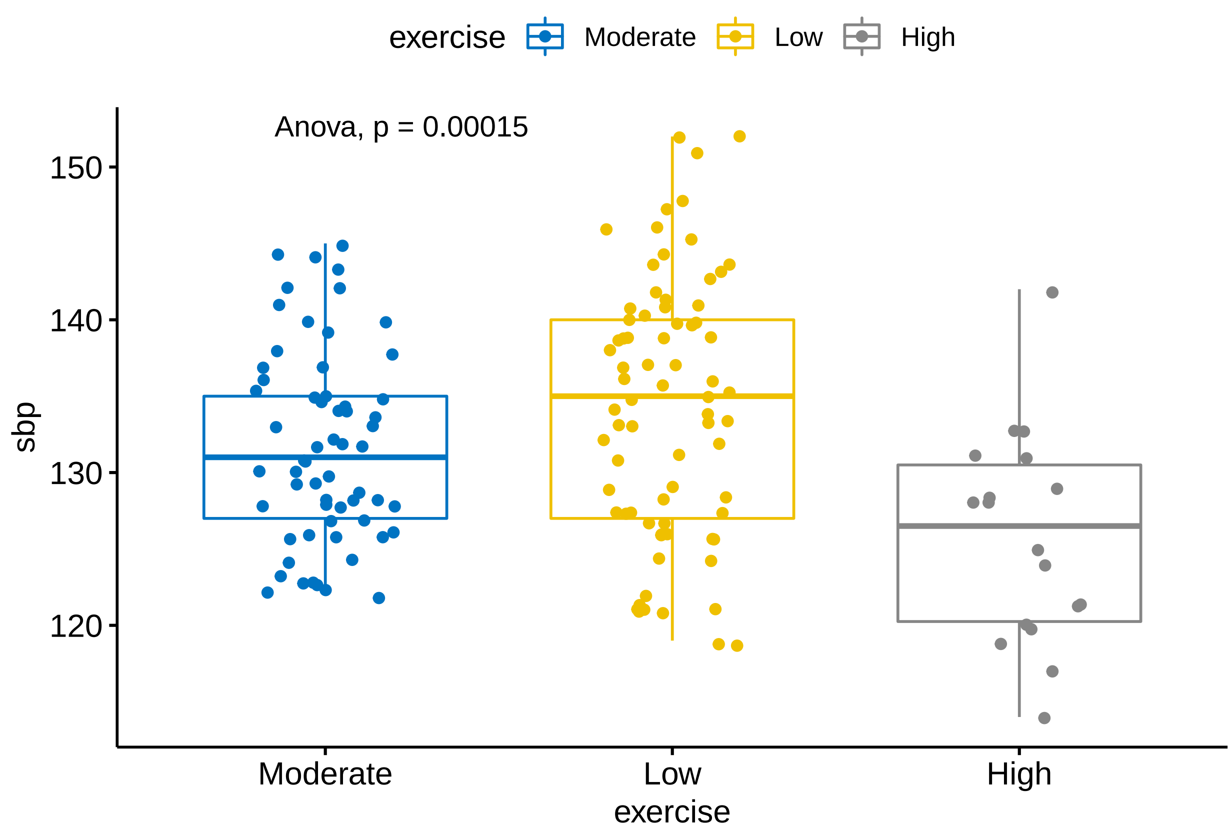
```
##
##  Descriptive statistics by group
## group: High
##    vars  n   mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 18 125.78 6.97  126.5   125.5 8.15 114 142    28 0.35    -0.48 1.64
## ------------------------------------------------------------
## group: Low
##    vars  n   mean   sd median trimmed   mad min max range skew kurtosis se
## X1    1 74 134.24 8.57    135  134.15 10.38 119 152    33 0.03    -0.92  1
## ------------------------------------------------------------
## group: Moderate
```

```
##    vars  n   mean   sd median trimmed  mad min max range skew kurtosis  se
## X1    1 61 131.74 6.28    131  131.45 5.93 122 145    23 0.33    -0.84 0.8
```

```
one.way =aov(sbp ~ exercise, data = hstat)
summary(one.way)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## exercise      2   1064   532.0   9.324 0.000152 ***
## Residuals   150   8559    57.1                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ggboxplot(hstat, x = "exercise", y = "sbp",
          color = "exercise",
          palette = "jco",
          add = "jitter") +
          stat_compare_means(method = "anova")
```




```
#yes, there is significant effect of exercise on sbp
#but, which pair comparison has most effect?

leveneTest(sbp ~ exercise, data = hstat, center=mean)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##        Df F value  Pr(>F)
## group   2  4.2458 0.01608 *
##       150
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#significant p value, thus equal variance not assumed
#use Welch one way ANOVA

oneway.test(sbp ~ exercise, data = hstat) #for overall significance
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  sbp and exercise
## F = 9.5994, num df = 2.000, denom df = 48.988, p-value = 0.0003037
```

```
pairwise.t.test(hstat$sbp, hstat$exercise,
                p.adjust.method = "BH", pool.sd = FALSE)
```

```
##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  hstat$sbp and hstat$exercise
##
##          High    Low
## Low      0.00035 -
## Moderate 0.00470 0.05254
##
## P value adjustment method: BH
```

```
#if equal variance assumed, use Tukey
tukey.one.way<-TukeyHSD(one.way)
tukey.one.way
```

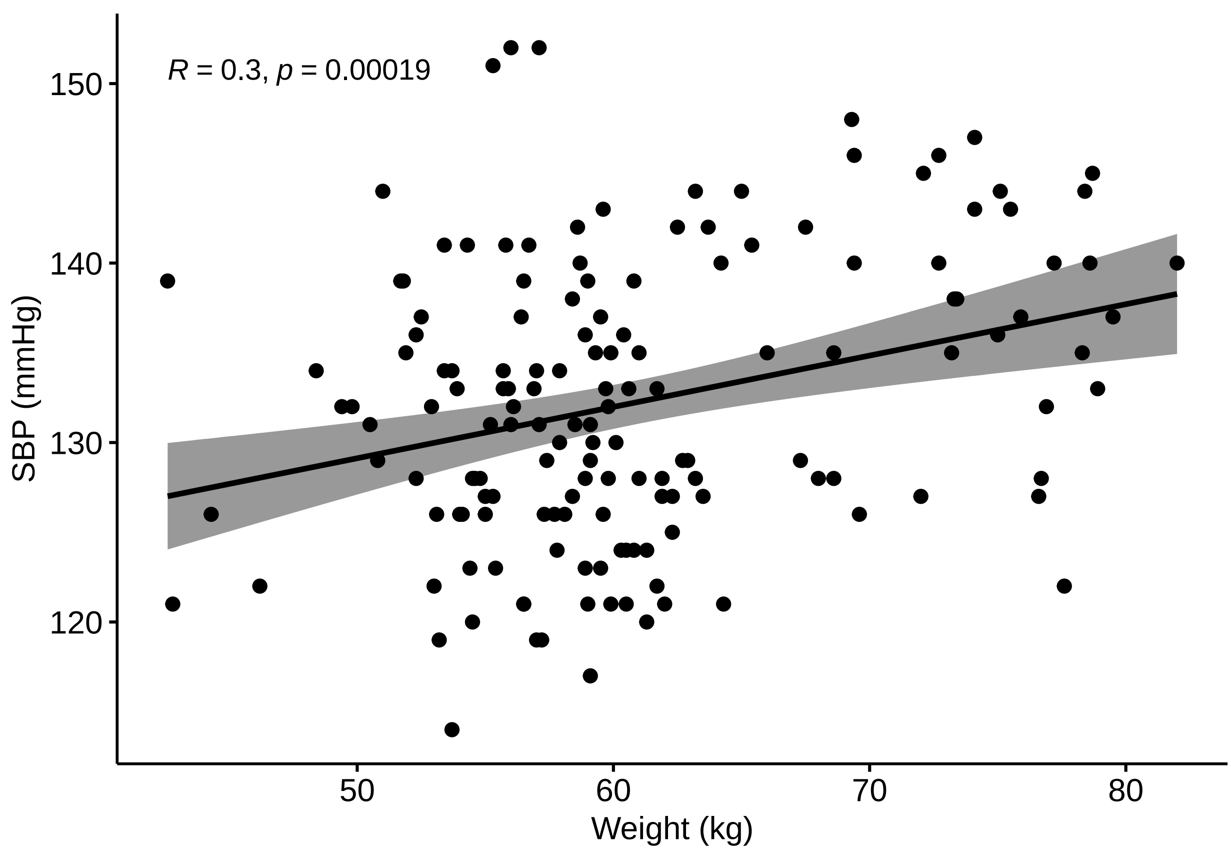
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = sbp ~ exercise, data = hstat)
##
## $exercise
##                    diff       lwr        upr     p adj
## Low-High       8.465465  3.766172 13.1647594 0.0001043
## Moderate-High  5.959927  1.163662 10.7561923 0.0105140
## Moderate-Low  -2.505538 -5.597805  0.5867282 0.1371181
```

```
#Exercise adds benefit in sbp reduction
#does it correlate with weight?

#pearson correlation coefficient test

cor.test(hstat$wt,hstat$sbp, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  hstat$wt and hstat$sbp
## t = 3.8267, df = 151, p-value = 0.0001897
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1455186 0.4354641
## sample estimates:
##       cor
## 0.2973312
```

```
ggscatter(hstat, x = "wt", y = "sbp",
          add = "reg.line",
          conf.int = TRUE,
          cor.coef = TRUE,
          cor.method = "pearson",
          xlab = "Weight (kg)", ylab = "SBP (mmHg)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



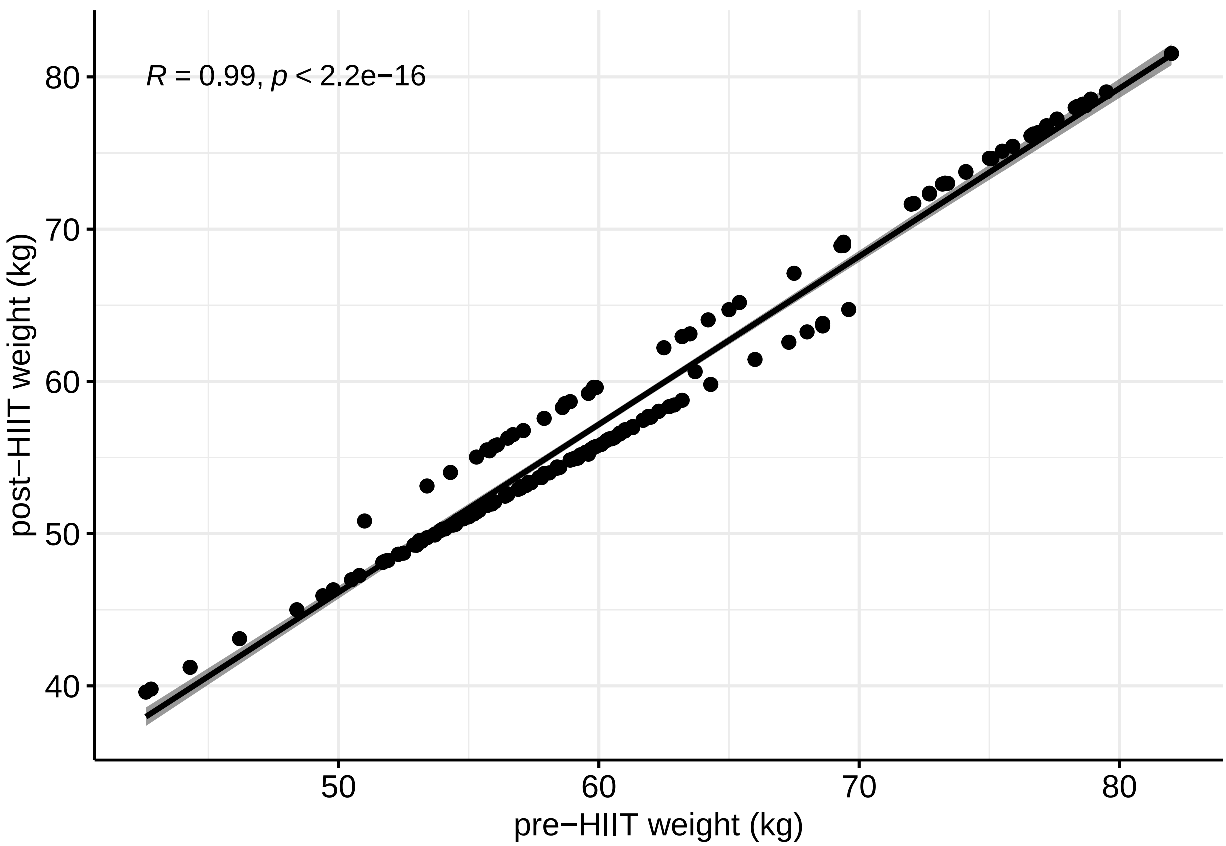

```
#yes, the heavier the person, the higher the sbp
#now you conducted a high-intensity interval training (HIIT) intervention
#you want to measure pre and post HIIT effect on weight
#paired t test
```

```r
t.test(hstat$wt, hstat$wt2, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  hstat$wt and hstat$wt2
## t = 19.015, df = 152, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.445299 3.012348
## sample estimates:
## mean of the differences
##                2.728824
```

```r
ggscatter(hstat, x = "wt", y = "wt2",
          add = "reg.line",
          conf.int = TRUE,
          cor.coef = TRUE,
          cor.method = "pearson",
          xlab = "pre-HIIT weight (kg)", ylab = "post-HIIT weight (kg)") +
          grids(linetype = "solid")
```

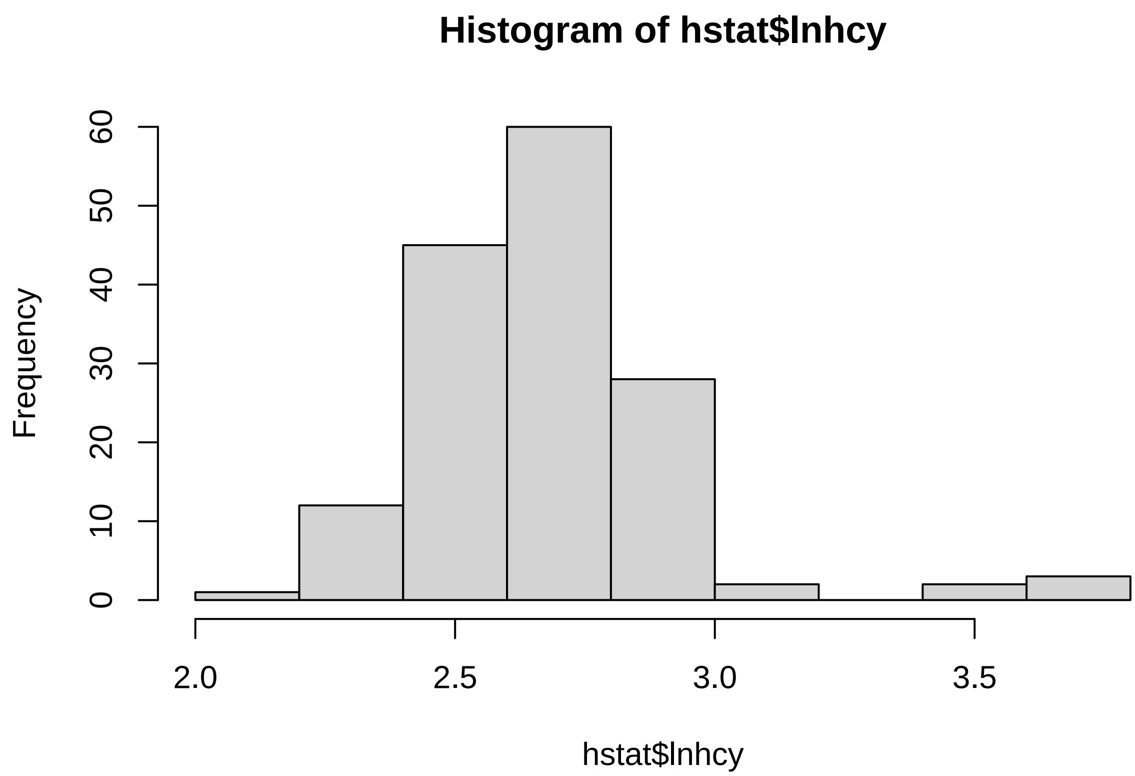
```
## `geom_smooth()` using formula 'y ~ x'
```



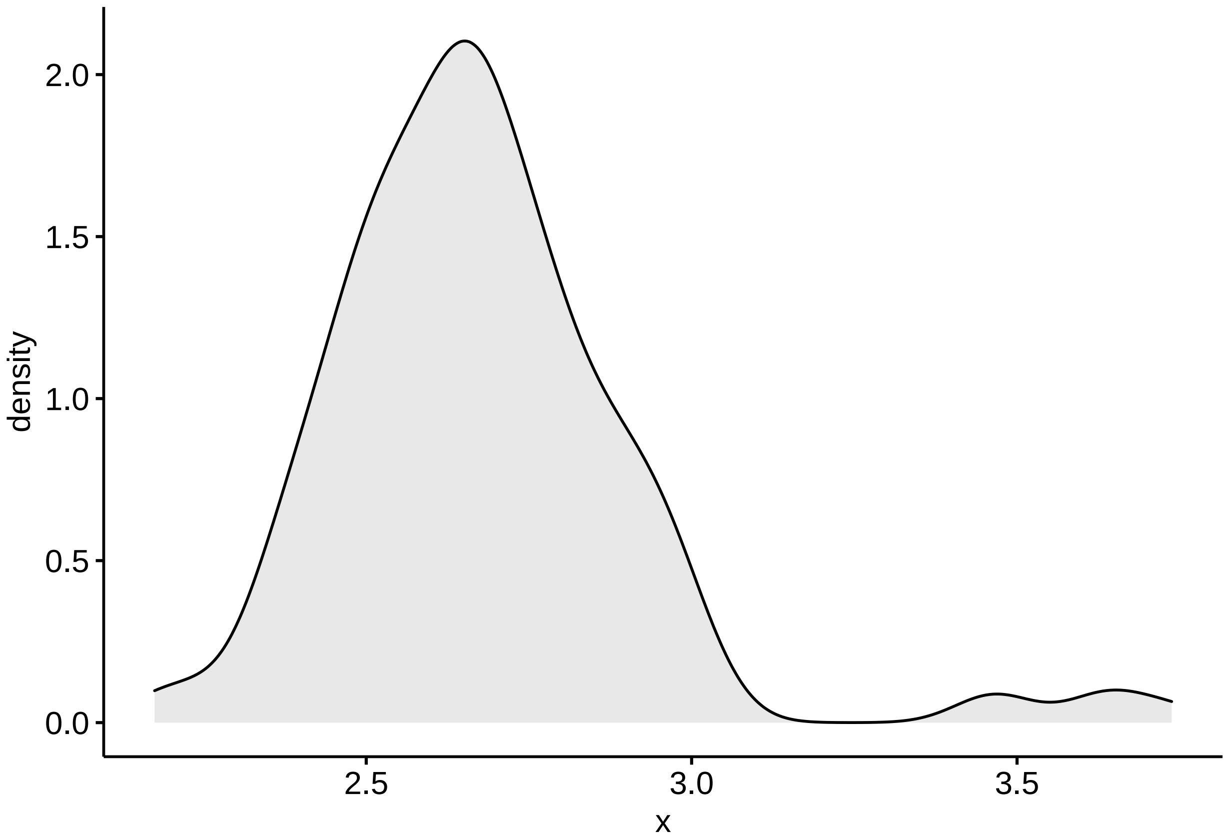
```r
#Now we know that sbp is affected by weight.
#exercise gives additional benefit to weight reduction
#you are now concerned with the exercise giving effect on cardiovascular health
#homocysteine (hcy) relates to cardiovascular heath from literature
#what are the factors contributing to hcy level?
```

```r
#comparing numerical values: non-parametric
```

```r
#try transforming data into normal distribution by ln
hstat$lnhcy= log(hstat$hcy)
hist(hstat$lnhcy)
```




```r
ggdensity(hstat$lnhcy, fill = "lightgray")
```




```r
#still not normally distributed
#need to do non-parametric test

#female has higher or lower hcy level compared to male?
#mann whitney U test

install.packages("SmartEDA")
```

```
## Warning in install.packages :
##   package 'SmartEDA' is in use and will not be installed
```

```r
library(SmartEDA)
ExpCustomStat(hstat,
              Cvar="sex",
              Nvar="hcy",
              stat=c("median","IQR"),
              gpby=TRUE,
              dcast=F)
```

```
##       sex Attribute median   IQR
## 1:   Male       hcy   14.3 3.650
## 2: Female       hcy   14.1 3.575
```

```
wilcox.test(hcy~sex, data=hstat) #synonym to Mann Whitney U test
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  hcy by sex
## W = 2658.5, p-value = 0.3675
## alternative hypothesis: true location shift is not equal to 0
```

```
#sex has no signififant relationship with hcy

#how about exercise intensity?
#kruskal wallis test


ExpCustomStat(hstat,
              Cvar="exercise",
              Nvar="hcy",
              stat=c("median","IQR"),
              gpby=TRUE,
              dcast=F)
```

```
##    exercise Attribute median   IQR
## 1: Moderate       hcy  14.30 2.600
## 2:      Low       hcy  14.55 5.075
## 3:     High       hcy  12.35 2.350
```

```
kruskal.test(hcy ~ exercise, data = hstat) #if significant, proceed with pairwise comparison
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  hcy by exercise
## Kruskal-Wallis chi-squared = 11.436, df = 2, p-value = 0.003286
```

```
pairwise.wilcox.test(hstat$hcy, hstat$exercise,p.adjust.method = "BH")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  hstat$hcy and hstat$exercise
##
##          High  Low
## Low      0.005 -
## Moderate 0.014 0.133
##
## P value adjustment method: BH
```

```
#high intensity exercise significantly gives lower HCY compared to low/mod
#you proceed in continuing the HIIT intervention as it gives benefit to both sbp and hcy
```

```r
#measure effectiveness again on weight reduction
#wilcoxon signed rank test

wilcox.test(hstat$wt,hstat$wt2,paired=TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  hstat$wt and hstat$wt2
## V = 11781, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```r
#yes, the HIIT is effective in weight reduction
#however you noticed some of your respondents are diabetic
#worry that your intervention gives more harm than good
#finding relationship between hba1c and both sbp/hcy
#spearman correlation coefficient test

cor.test(hstat$hba1c,hstat$sbp, method="spearman")
```

```
## Warning in cor.test.default(hstat$hba1c, hstat$sbp, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  hstat$hba1c and hstat$sbp
## S = 416312, p-value = 0.0001441
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3025482
```

```r
cor.test(hstat$hba1c,hstat$hcy, method="spearman")
```

```
## Warning in cor.test.default(hstat$hba1c, hstat$hcy, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  hstat$hba1c and hstat$hcy
## S = 515097, p-value = 0.09116
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1370514
```

```
#you conclude that only sbp has a significant correlation with hba1c
#in future, you would prioritize
#giving HITT intervention to the diabetic & hypertensive patients

#now, you are focused back to your objective 3
#factors contributing to hypertension (hpt)
```

```
#comparing categorical variables
```

```
install.packages("dplyr")
```

```
## Warning in install.packages :
##   package 'dplyr' is in use and will not be installed
```

```
library(dplyr)

hstat2<-hstat %>%
  mutate(hpt=if_else(hstat$sbp<140 & hstat$dbp<90,'normal','high'))
View(hstat2)

#smoking has a relationship with hpt?
#chi square test

chisq.test(hstat2$hpt,hstat2$smoking,correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  hstat2$hpt and hstat2$smoking
## X-squared = 15.607, df = 1, p-value = 7.797e-05
```

```
chisq.test(hstat2$hpt,hstat2$smoking)$observed
```

```
##            hstat2$smoking
## hstat2$hpt No Yes
##     high   23  36
##     normal 67  27
```

```
#yes, smoking is significantly related to hpt. More smokers are hypertensive

#how about BMI status and hpt?

hstat3<- hstat2 %>%
  mutate(height_m = ht / 100,bmi = wt / (height_m^2))
View(hstat3)
hstat3$bmistatus<- cut(hstat3$bmi,
                                 breaks=c(-Inf, 18.49999, 24.9999, 29.9999, Inf),
                                 labels=c("underweight", "normal", "overweight", "obese"))

#fisher's exact test (used when more than 20% celss with expected count less than 5)
chisq.test(hstat3$hpt,hstat3$bmistatus)$expected
```

```
## Warning in chisq.test(hstat3$hpt, hstat3$bmistatus): Chi-squared approximation
## may be incorrect

##            hstat3$bmistatus
## hstat3$hpt underweight   normal overweight     obese
##     high      2.313725 28.92157    18.5098  9.254902
##     normal    3.686275 46.07843    29.4902 14.745098
```

```
fisher.test(hstat3$hpt,hstat3$bmistatus)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  hstat3$hpt and hstat3$bmistatus
## p-value = 1.205e-05
## alternative hypothesis: two.sided
```

```
#significant relationship between hpt and bmi status
```

```
#reporting your findings in table form
```

```
#package needed
#"sjPlot"
#"apaTables"

install.packages("sjPlot")
```

```
## Error in install.packages : Updating loaded packages
```

```
library(sjPlot)
install.packages("apaTables")
```

```
## Error in install.packages : Updating loaded packages
```

```
library(apaTables)

#table created in word file in your directory!
#sjt.xtab(hstat3$smoking, hstat3$hpt, file = "sjt_contingency.doc")
#apa.aov.table(one.way, filename="Table_anova.doc", table.number = 2)
```